ELSEVIER

Cognition 86 (2002) 123–155

COGNITION

www.elsevier.com/locate/cognit

Why do children learn to say "Broke"? A model of learning the past tense without feedback

Niels A. Taatgen[a,*], John R. Anderson[b]

[a]*Department of Artificial Intelligence, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands*
[b]*Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA*

Received 20 May 2000; received in revised form 26 February 2002; accepted 20 August 2002

Abstract

Learning the English past tense is characterized by a U-shaped learning function for the irregular verbs. Existing cognitive models often rely on a sudden increase in vocabulary, a high token-frequency of regular verbs, and complicated schemes of feedback in order to model this phenomenon. All these assumptions are at odds with empirical data. In this paper a hybrid ACT-R model is presented that shows U-shaped learning without direct feedback, changes in vocabulary, or unrealistically high rates of regular verbs. The model is capable of learning the default rule, even if regular forms are infrequent. It can also help explore the question of why there is a distinction between regular and irregular verbs in the first place, by examining the costs and benefits of both types of verbs. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Cognitive development; Language; Acquisition; Cognitive models; Past tense

1. Introduction

Learning the past tense has been the subject of debate in cognitive science since Rumelhart and McClelland (1986) first modeled it as part of their parallel distributed processing effort. Numerous authors have contributed to the issue since, criticizing the original model (e.g. Pinker & Prince, 1988) or offering alternatives, either connectionist (MacWhinney & Leinbach, 1991; Plunkett & Juola, 1999; Plunkett & Marchman, 1991, 1993) or symbolic (Ling & Marinov, 1993). Although each of these models offers contributions to the debate, they leave some issues unaddressed, and sometimes make assumptions that are not entirely realistic. Fortunately, more empirical data have become

* Corresponding author.
E-mail address: niels@ai.rug.nl (N.A. Taatgen).

0010-0277/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0010-0277(02)00176-2

available on the topic, mainly through a detailed review of the available data by Marcus et al. (1992).

One of the main topics in learning the past tense is U-shaped learning. Traditionally, three stages are distinguished. In the first stage, when the child starts using past tenses, irregular verbs are used correctly. In the second stage, the child develops a sense for the regularity in regular past tenses. As a consequence, it will now sometimes construct past tenses of irregular verbs in a regular way (e.g. *go-goed* as opposed to *go-went*). In the third stage, this overregularization diminishes until performance is without errors. Since performance on irregular verbs is worst in the second stage, the performance curve has a U-shape, hence the name of the phenomenon. The interesting question is what causes this U-shape.

One account focuses on a *dual-representation* of knowledge: on the one hand past tenses are memorized as separate cases and on the other hand a rule is learned that can produce regular past tenses (Marcus et al., 1992; Pinker & Prince, 1988). According to the dual-representation explanation, in the first stage only separate cases are memorized. This means that in the first stage producing past tenses is only partially successful, because if a past tense has not been memorized it cannot be reproduced. This changes in the second stage, because at that moment the regular rule is learned. The regular rule can produce a past tense for any verb, although this may be an incorrect one. These incorrect past tenses, overgeneralizations, slowly disappear as more correct examples are learned, and a gradual transition to stage 3 is made. An important aspect of the theory, the *blocking mechanism*, states that the regular rule is applied unless an exception can be retrieved from memory. By assuming that the process of memory retrieval is noisy and occasionally fails to retrieve an irregular past tense, the U-shape can be explained. To summarize: in stage 1, *broke* is produced when retrieval is successful, and unsuccessful retrievals go undetected. In stage 2, successful retrievals still produce *broke*, but unsuccessful retrievals now result in the application of the rule, producing *breaked*. In stage 3 the memory trace for *broke* is strong enough to always block the rule.

The dual-representation explanation leaves a number of questions. How is the regular rule learned? Why is retrieving of examples the dominant strategy, so that it can block the application of rules? This is certainly not a general cognitive principle, as in many cases the cognitive systems strives for generalization.

The *single-representation explanation* only uses a single representational system, usually a neural network, to explain past-tense learning. In this explanation, U-shaped learning is mainly initiated by changes in vocabulary size. When the vocabulary is still small, the network is large enough to accommodate separate cases, but as the vocabulary grows the need for regularization increases. At some point during the learning, the network shifts its weights to support regularization, but needs some time to properly integrate this with the exceptions, causing the U-shape. The single-representation explanation has a number of problems as well. In order to get the desired behavior, certain assumptions have to be made about the input, more specifically about the growth of the vocabulary. Unfortunately, there is little evidence for these assumptions in actual data, requiring an additional assumption about the distinction between *input* (the raw input from the environment) and *uptake* (what the child actually processes). A final issue is the problem of feedback. Children generally receive no feedback on the syntactic correctness of the

language they produce, but most network models need the correct answer in order to adjust their weights. Additional assumptions are needed to explain this source of information.

The model that we will present here is based on dual-representation theory. It offers a mechanism that enables the model to learn the regular rule. Also, it is able to learn that blocking is a good strategy in past-tense generation, which is usually just assumed in dual-representation theory.

1.1. What are the facts that need to be explained?

In the remainder of the article we will examine *models* of learning the past tense. A model is an instantiation of the theory that actually goes through the process of learning the past tense, producing learning behavior that can be tested against empirical facts. The main fact we are focusing on is U-shaped learning. The main criterion for a U-shape is specified in the three stages outlined above. Researchers have however identified a number of additional aspects related to the U-shape that serve as important criteria to test theories.

A first aspect of the regular rule in English is that it is a default rule. This is characterized by the fact that, given an unknown verb, people tend to use a regular past tense. Also, words from foreign languages are regularized, as are denominal verbs, nouns that are used as verbs (Marcus, Brinkman, Clahsen, Wiese, and Pinker (1995) list 21 different circumstances in which the default rule is applied).

A second aspect of the regular rule is the role of frequency. In the English past tense, regular verbs have a high *type-frequency*: most of the verbs are regular. However, although there are only few irregular verbs, they tend to be used very often. As a consequence, the *token-frequency* of regular verbs, how often regular verbs occur in spoken or written speech, is actually much lower. Despite the fact that there are many more regular verbs (type-frequency), their occurrence (token-frequency) is about 25–30% of all verbs used. A model of the past tense should be able to learn the past tense, and exhibit U-shape learning behavior, given these input distributions.

The English past tense is not the only example of inflection in which there is a default rule and a set of exceptions. A model of the past tense should be able to account for other cases as well. An interesting case is the German plural (Clahsen, Rothweiler, Woest, & Marcus, 1992). The regular German plural has both a low type-frequency (no more than 10%) and a low token-frequency (no more than 5%). Despite these low frequencies, it has many properties of a default rule (Marcus et al., 1995). Although this view has not remained unchallenged (Hahn & Nakisa, 2000; Plunkett & Nakisa, 1997), it stresses the need for models of regular and irregular inflection to be able to learn default rules based on low frequencies.

A third aspect is a property of language acquisition in general: parents generally do not give children adequate feedback regarding syntactical errors (Pinker, 1984). Children therefore perceive correct past tenses as they are used by their parents and others, but do not receive feedback on their own production. This aspect is important for an explanation of the transition from overgeneralization to correct behavior. In general, repetition reinforces behavior, so why does a child instead abandon a certain behavior (overgeneralization), if it receives no feedback on its incorrectness?

A fourth issue is the question whether or not the onset of overregularization is driven by

the growth of the vocabulary, in the sense that a certain "critical mass" is needed before generalization becomes worthwhile, or whether overregularization is independent of vocabulary. The strongest indication for the critical mass hypothesis would be when a sudden increase in vocabulary size coincides with the onset of overregularization. Marcus et al. (Marcus, 1996; Marcus et al., 1992) found no such coincidences in their data. Marchman and Bates (1994), in another empirical study, didn't find such coincidences either, as they just found a linear increase in vocabulary size. They did however find a slightly weaker indication for the critical mass hypothesis. They found a non-linear relation between the size of the vocabulary and the number of irregular verbs that were inflected for the past tense, both correct and overregularized: the number of types inflected grows faster as the vocabulary becomes larger. The fact that the relation is non-linear can be interpreted as some support for the critical mass hypothesis.

A final issue is the question of why there is a distinction between irregular and regular verbs in the first place. Why are not all verbs either regular or irregular? Perhaps this has something to do with the fact that the child has to learn to use the past tense without getting feedback on its own production. As a consequence, it has to invent its own grammar based on positive examples. Maybe there is something about language use that encourages the development of regular and irregular verbs which would occur even in the total absence of any feedback (Hare and Elman (1995) investigate this issue as well, using a slightly different method).

1.2. Overview of the Adaptive Control of Thought, Rational (ACT-R) model

The model we will present here uses several strategies to produce past tenses. The model starts out with three strategies: retrieval, analogy and the zero strategy, and later learns a fourth strategy, the regular rule. The retrieval strategy tries to produce a past tense by recalling an example of inflecting the word from memory. The successfulness of this strategy depends on the availability of examples. The analogy strategy recalls an arbitrary example of a past tense from memory, and tries to use this as a basis for analogy. Analogy will only succeed if it can find a pattern in the example that is applicable to the current word. The zero strategy (or do-nothing strategy) always succeeds, as it does not attempt any inflection at all. Associated with each strategy is the expected outcome of that strategy. The estimate for expected outcome is continuously updated on the basis of experiences of how much effort it takes to use a strategy. The strategy with the highest expected outcome has the highest probability of being tried first, and if it fails other strategies can be attempted. The fourth strategy, the regular rule, is learned on the basis of the analogy strategy. By substituting an example of regular inflection into the rules that implement analogy, the regular rule is obtained. It takes some time for the regular rule to surface because new rules can only be learned when the parent rules have sufficient experience and because the new rule starts out with a relatively low expected outcome and first has to prove itself.

The U-shaped learning can be explained by the dynamics of the expected outcomes of the different strategies, the introduction of the regular-rule strategy, and the increased availability of examples of past tenses in memory. The assumption of the model is that it both perceives (correct) past tenses in the environment, and produces them itself. The

model receives no feedback on its own production, except internal feedback on the effort associated with producing it. During stage 1 the main strategies are retrieval and do-nothing. Retrieval, if it succeeds, will generally produce correct irregular verbs, while do-nothing produces undetectable errors that are not counted. Performance on irregular verbs will therefore be near 100% correct. Shortly after the regular rule is learned, the transition to stage 2 begins. An important aspect of the model is that the regular rule will not dominate the retrieval strategy, because irregular past tenses have certain advantages over regular past tenses. Although the expected outcome of the regular rule will remain lower than the retrieval strategy, whenever a retrieval fails to find an example the regular rule is applied. This produces overregularization errors in irregular verbs, especially in low-frequency verbs. As the model learns more and more correct past tenses (and occasionally incorrect past tenses due to its own production errors), the retrieval strategy eventually completely dominates the regular-rule strategy, resulting in stage 3 performance.

The mechanisms that produce the model's behavior are part of the *ACT-R cognitive architecture* (Anderson & Lebiere, 1998) that have been separately developed and tested in non-linguistic research.

1.3. An overview of some of the existing models

The first model of learning the past tense was by Rumelhart and McClelland (1986). It consisted of a two layer feed-forward network that performed surprisingly well on the task. Although it produced U-shaped behavior, it did have some severe problems. One of the main criticisms (see also Pinker & Prince, 1988) is that the onset of the U-shape corresponds to a change in vocabulary from ten to 420 verbs, a sudden jump in vocabulary that cannot be justified. MacWhinney and Leinbach (1991) tried to address many of the criticisms of Pinker and Prince using a three-layer back-propagation model, but lost part of the U-shaped learning in the process. Their model wasn't able to exhibit stage 1 behavior: correct behavior before the onset of overregularization. Plunkett and Marchman (1991) introduced the notion of *micro U-shaped learning*, the fact that overregularization is a phenomenon at the level of individual words instead of being attributable to global stage-wise development. Although their model, also a three-layered network, produced such micro U-shapes, it did not produce global U-shaped learning. Their later update of the model (Plunkett & Marchman, 1993) did produce this global U-shape. One of the problems for the network models up to this model was to reach the initial stage of error-free production of irregulars, before the onset of overregularization. Plunkett and Marchman solved this problem by training the network on the initial 20 verbs until it was 100% correct. After that they started to expand the vocabulary gradually. When the size of the vocabulary was about 100 verbs, performance on irregular verbs started to degrade, and overgeneralization commenced. Around size 200 the network recovered, and at size 300 performed again at 100% accuracy. Although the original model introduces a discontinuity in the growth of the vocabulary at size 100, Plunkett and Marchman (1996) later showed that their model was able to capture U-shaped learning without this discontinuity as well. One might criticize the model by pointing out that it creates stage 1 behavior artificially, exactly the stage previous models had had problems dealing with. This is also

reflected in the fact that the model exhibits U-shaped learning in the regular verbs as well: when performance on irregular verbs decreases, the performance on regular verbs also decreases. This is rather unexpected, and not backed up by data, as Marcus et al. (1992) found that decreases in performance on irregular verbs coincide with *increases* in performance on regular verbs. A more recent model by Plunkett and Juola (1999) has similar problems. The input scheme they use is similar to Plunkett and Marchman (1993): an initial vocabulary of 20 words is trained until performance is perfect, after which the vocabulary is increased. But instead of using the previous, discontinuous training scheme, they increase it exponentially. This produces U-shaped learning in the brief period that the model moves from around 20% of its vocabulary to 100%. Only after the vocabulary has hit the ceiling at 100% does performance start to recover. As in the previous model, the U-shape in irregular verbs coincides with a U-shape in regular verbs.

Typically in neural network models the input is presented in the form of epochs consisting of the set of verbs the model has to inflect. As described above, this set is gradually expanded over time. The use of these epochs has the problem that high-frequency words are presented as often as low-frequency words. Most models mend this problem by inserting multiple copies of high-frequency words, but hardly ever in the proportion that reflects the real world (as plotted for 478 verbs in Fig. 1).

In summary, the behavior of neural network models of the past tense depends heavily on the structure of the input, both in growth and constitution. This is also a basis for criticism: as Marcus (1995) points out, the structure of the input the networks receive in no way mirrors the input children receive or the output they produce. The token-frequency of

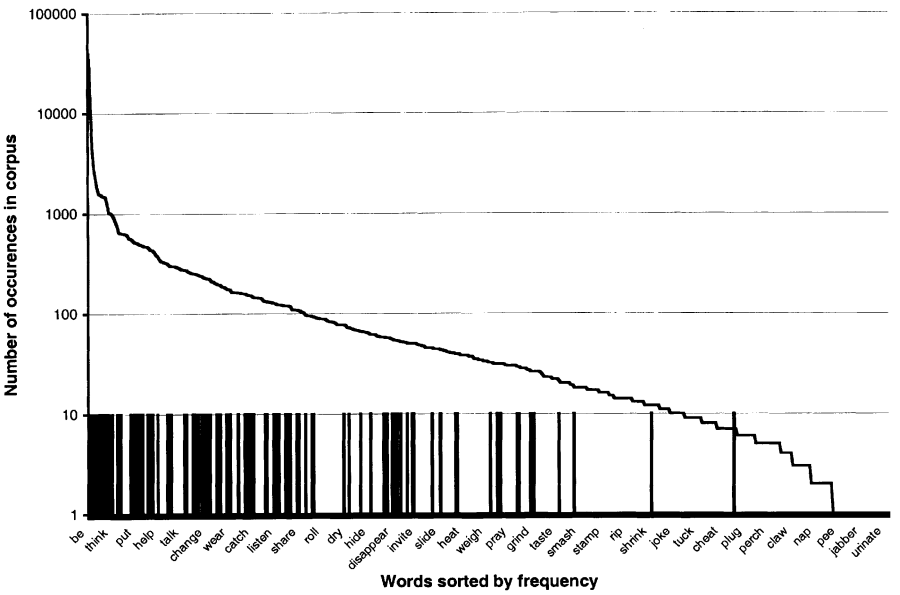

Fig. 1. Frequencies of 478 words used by children or their parents from Marcus et al. (1992) according to the Francis and Kucera (1982) corpus. The curve denotes the number of occurrences in the corpus while the bars indicate irregular verbs.

regular verbs is rather low, in the order of 25–30%. Network models often require a high frequency of regular verbs to capture the regularity (at least models of the English past tense that also reproduce U-shaped learning). For example, Plunkett and Marchman (1993) note:

> (…) generalization is virtually absent when regulars contribute less than 50% of the items overall. (…) In summary, the level of generalization observed in these simulations is closely related to the total number of regular verbs in the vocabulary, provided the proportion of regulars exceeds the 50% level. (p. 55)

Plunkett and Marchman (1996) counter some of Marcus' criticism by showing in several runs of their model that the discontinuity in the input does not necessarily coincide with the onset of overregularization, and that even without a discontinuity U-shaped learning still occurs. It is therefore odd that Plunkett and Juola (1999) do not use a linearly increasing vocabulary size, but an implausible exponential growth.

Most criticisms concerning the input are addressed by distinguishing between the *input* of the child and the *uptake* of the child (Plunkett & Marchman, 1993). The input is what the child actually perceives and has to produce, and the uptake is some subset of this input that is fed into the network. They defend this by pointing out that low-frequency words need to be processed sufficiently. But this defense is based on the properties of neural networks, and not on properties of human learning. Neural networks suffer from "catastrophic forgetting" if certain items do not appear often enough in the input. This problem has never been demonstrated in humans. The input/uptake problem becomes more salient if one wants to model the German plural, where the type-frequency for regulars is at most 10% and the token-frequency at most 5%.

Another issue is feedback. All network models need feedback on their output in order to be able to adjust their weights. As children do not receive feedback on their output, a different explanation of feedback has to be given. Plunkett and Juola (1999) offer the following explanation:

> The child is continually taking in word tokens and comparing the words actually heard (e.g. "went") to the tokens that the child's hypothesis generator would have expected to produce as inflected forms of a given stem; when they differ, this provides evidence to the child that the hypotheses are wrong and should be modified. (p. 466)

So in order to learn the past tense, the child has to hear a past tense from a parent or other speaker, and has to determine what the stem is. Consequently the stem has to be selected for "uptake", and if it is, it is fed into the network to determine the past tense again. Its performance is then compared to the past tense initially heard, and weights are adjusted accordingly. When the child actually has to produce a past tense, the network is used without any learning, as there is no feedback to adjust its weights. This implies language production itself has no impact at all on performance, defying the general idea that practice is an important aspect of learning.

Some recent connectionist models have been made of both the Arabic plural (Plunkett & Nakisa, 1997) and the German plural (Hahn & Nakisa, 2000). These models do not attempt to exhibit U-shaped learning, but otherwise have some interesting characteristics. In the

Hahn and Nakisa (2000) model, the network is trained on a partial vocabulary and is tested on the rest of it. The model did very well on unknown words, being correct 81% of the time. This indicates that information is contained within the phonological structure of the word, enabling the model to often guess the right inflection correctly. Nevertheless the model does not learn to apply the *-s* default rule. Instead, Hahn and Nakisa challenge the Marcus et al. (1995) claim that German has a default rule at all. The fact remains though, that German speakers use the *-s* suffix much more often than their model.

2. The ACT-R architecture

The basic theoretical foundation of the ACT-R architecture is *rational analysis* (Anderson, 1990). According to rational analysis, each component of the cognitive system is optimized with respect to demands from the environment, given its computational limitations. The main components in ACT-R are a declarative (fact) memory and a production (rule) memory. To avoid confusion with grammatical rules, we will refer to rules in production memory with *production rules*. ACT-R is a so-called hybrid architecture, in the sense that it has both symbolic and sub-symbolic aspects. We will introduce these components informally. Table 1 provides a formal specification of some critical aspects of

Table 1
ACT-R equations[a]

Equation	Description
Activation	
$A = B + \text{context} + \text{noise}$	The activation of a chunk has three parts: base-level activation, spreading activation from the current context and noise. Since spreading activation is a constant factor in the models discussed, we treat activation as if it were just base-level activation.
Base-level activation	
$B(t) = \log \sum_{j=1}^{n} (t - t_j)^{-d}$	n is the number of times a chunk has been retrieved from memory, and t_j represents the time at which each of these retrievals took place. So, the longer ago a retrieval was, the less it contributes to the activation. d is a fixed ACT-R parameter that represents the decay of base-level activation in declarative memory.
Retrieval time	
$\text{Time} = Fe^{-fA}$	Activation determines the time required to retrieve a chunk. A is the activation of the chunk that has to be retrieved, and F and f are fixed ACT-R parameters. Retrieval will only succeed as long as the activation is larger than retrieval threshold τ, which is also a fixed parameter.
Expected outcome	
$\text{Expected outcome} = P_{\text{p}}G - C_{\text{p}} + \text{noise}$	Expected outcome is based on three quantities, the estimated probability of success of a production rule (P), the estimated cost of the production rule (C), and the value of the goal (G).

[a] These equations are simplified versions of the original Anderson and Lebiere (1998) equations.

the subsymbolic level. Further details about the architecture can be found in Anderson and Lebiere (1998).

Items in declarative memory, called *chunks*, have different levels of *activation* to reflect their use: chunks that have been used recently or chunks that are used very often receive a high activation. This activation decays over time if the chunk is not used. Activation represents the probability (actually, the log odds) that a chunk is needed and the estimates provided for by ACT-R's learning equations represent the probabilities in the environment very well (see Anderson, 1993, Chap. 4, for examples). The level of activation has a number of effects. One effect of activation is that when ACT-R can choose between chunks, it will retrieve the chunk with the highest activation. Activation also affects retrieval time. As the activation of a chunk decreases, its retrieval time grows exponentially. At some point it is no longer feasible to retrieve a chunk: it would just take too much time. Because of this ACT-R is not able to retrieve chunks with an activation below a certain threshold.

Chunks cannot act by themselves, they need *production rules* for their application. In order to use a chunk, a production rule has to be invoked that retrieves it from declarative memory and does something with it. Since ACT-R is a goal-driven theory, chunks are always retrieved to achieve some sort of goal. In the context of learning the past tense the goal is simple: given the stem of a word, produce the past tense. One strategy to produce a past tense is to just retrieve it from memory, using a production rule like:

```
IF    the goal is to produce a past tense of a word
      AND there is a chunk that specifies the past tense of that
      word
THEN  set the answer of the goal to the past tense
```

If the goal is to produce a past tense of a certain word, this production rule will attempt to retrieve a chunk from declarative memory that specifies what the past tense is. Of course this production rule will only be successful if such a fact is present and its activation is high enough.

The behavior of production rules is also governed by the principle of rational analysis. Each production rule has a real-value quantity associated with its expected outcome. This expected outcome is calculated from estimates of the cost and probability of reaching the goal if that production rule is chosen. The unit of cost in ACT-R is time. ACT-R's learning mechanisms constantly update these estimates based on experience. If multiple production rules are applicable for a certain goal, the production rule is selected with the highest expected outcome.

In both declarative and procedural memory, selections are made on the basis of some evaluation, either activation or expected outcome. This selection process is noisy, so the item with the highest value has the greatest probability of being selected, but other items get opportunities as well. This may produce errors or suboptimal behavior, but also allows the system to explore knowledge and strategies that are still evolving.

In addition to the learning mechanisms that update activation and expected outcome, ACT-R can also learn new chunks and production rules. New chunks are learned auto-

matically: each time a goal is completed it is added to declarative memory. If an identical chunk is already present in memory, both chunks are merged and their activation values are combined. New production rules are learned on the basis of specializing and merging existing production rules. The circumstance for learning a new production rule is that two rules fire one after another with the first rule retrieving a chunk from memory. A new production rule is formed that combines the two into a macro-rule but eliminates the retrieval. The macro-rule is specialized to contain that information that was retrieved. This process is reasonably safe, in the sense that it never produces rules that are completely different from the production rules already present, but can nevertheless produce radical changes in behavior if the new rule outperforms the old rule by a large measure.[1]

3. A rational account of regular and irregular past tense

Why is there a distinction between regular and irregular verbs in the first place? Both the regular and the irregular forms of past tense represent possible ways of modifying a present tense verb to mark its tense. The regular rule adds an extra morpheme to the stem. In the case of irregular verbs the stem itself is changed (except for cases like hit-hit, where the past tense is identical to the present tense). Using a regular rule seems a very economical solution from the viewpoint of memory: only one rule is needed to produce past tenses for all verbs. Since there is no systematic way in which the stem itself can be changed, irregular verbs have to be memorized separately. What is the advantage of using an irregular past tense? Why not have a single rule for the past tense and be done with it? There are several possible reasons. First, a regular past tense is usually slightly longer than an irregular past tense. A regular rule always adds a morpheme to the stem that sometimes has to be pronounced as a separate syllable. A second reason why irregulars may have an advantage is that they are actually more regular from a phonetic viewpoint (Burzio, 1999). For example, **keeped* is phonetically irregular (in English) as opposed to *kept*. This phonetic disadvantage suggests that the use of a regular rule requires some phonetic post-processing that makes it less attractive than just storing and retrieving an irregular form. More generally, adding morphemes to stems of words lengthens them and may also distort the basic template of a word (Prasada & Pinker, 1993).

An alternative for inflecting a verb for past tense is to use no inflection at all. In that case tense has to be marked in a different way, for example by adding extra words like *yesterday*. An extra word has considerable costs, since it has to be selected and pronounced.

In this micro-economy of knowledge, the optimal choice of strategy depends on the frequency in which a word is used. High-frequency words benefit more from the irregular strategy, because the cases memorized turn up quite often. For low-frequency words the use of a rule is more optimal, since maintaining a case in memory for the few occasions the word is used does not overcome the disadvantage of using a rule. In ACT-R, this trade-off is already built into the basic mechanisms of the architecture. Due to activation learning (see the base-level activation equation in Table 1) low-frequency knowledge receives

[1] The process of proceduralization used in this model is not part of ACT-R 4.0, the current version of ACT-R, but is part of a proposal for the next version of the architecture.

lower activation than high-frequency knowledge. This activation difference translates into retrieval time and success: low-frequency items take more time to retrieve or cannot be retrieved at all.

Fig. 1 illustrates that this is the case: the 478 verbs (89 irregular, 389 regular) that children or their parents use, reported in Marcus et al. (1992), are sorted with respect to their frequency according to Francis and Kucera (1982). The curve shows the number of occurrences in the Francis and Kucera corpus, while a bar indicates an irregular verb. As can be seen in the graph, most irregular verbs are high-frequency words: the first regular verb is no. 13 (use). According to this distribution, only 25% of the words used (the token-frequency) are regular, which is close to the 30% Marcus et al. (1992) found in children's speech.

Different languages may strike different compromises between rule and example: in the German plural, for example, the default rule is quite rare (but default words are low-frequency). In other cases, for example the English gerund, produced by adding the *-ing* suffix to the stem, the rule fully dominates inflection.

In terms of rational analysis, U-shaped learning can be explained by a temporary imbalance between retrieving examples and using the rule, at a moment that the learning of the examples hasn't properly settled on the eventual activation values of the examples. This also explains why no feedback on performance is necessary: it is not the case that the cognitive system discovers that the regular rule is an overgeneralization, it is just that it hasn't properly memorized all the exceptions yet. This explanation closely matches the Marcus et al. (1992) explanation, except that they simply assume that the blocking system is the dominant strategy, while in our explanation the dominance of the irregular is a consequence of its greater efficiency.

In summary, the rational-analysis theory of ACT-R predicts that even in a situation where the cognitive system can choose between maintaining distinct past-tense forms (irregular past tenses) and adding a suffix to a word (regular past tenses), it will end up with high-frequency irregular verbs and low-frequency regular verbs. This will be the basis for the model.

4. A model of learning the past tense

In the previous section we mentioned several reasons why using a regular rule might have disadvantages. In the model we will adapt the phonetic post-processing explanation. This means that each time a past tense is constructed by adding a suffix to a stem, there are some extra costs involved in phonetic post-processing. The model initially has to choose between a number of ways to produce a past tense given the stem of the verb. Each of these methods is not specific to the task of producing a past tense, they are not even specific to language.

- Attempt to retrieve the past tense from declarative memory. Retrieving past cases from memory is a strategy that is used in almost any ACT-R model, and corresponds with the Logan (1988) instance theory of skill acquisition. (We will refer to this strategy as the *retrieve strategy*, or simply *retrieval*.)

- Attempt to generate a new past tense by analogy: retrieve an arbitrary past tense from memory and use it as a template to find a past tense for the current word. Analogy is also a strategy that is often used in ACT-R (e.g. Lebiere, Wallach, & Taatgen, 1998; Salvucci & Anderson, 1998) and is probably one of the dominant human strategies for problem solving and discovery. (We will refer to this strategy as the *analogy strategy*, or simply *analogy*.)
- Just use the stem as past tense, basically doing nothing at all. (We will refer to this as the *zero strategy* or *zero rule*.)

None of these strategies are very good initially. Analogy involves more than one reasoning step and is only successful if a suitable example is retrieved. The retrieve strategy needs examples before it can be successful. The zero rule always succeeds, but does not produce a past tense that can be distinguished from the present tense. Before the model can do anything useful beyond producing a past tense that is identical to the stem, it has to perceive some examples in the environment. Note that there is no production rule for the regular rule yet, ACT-R will learn it later on as a specialization of the analogy strategy. These initial strategies are similar to those proposed by MacWhinney (1978), who also suggested that the regular rule is formed on the basis of analogy.

4.1. A detailed description of the model

The model uses declarative-memory chunks to represent past tenses, both as a goal and as examples. A goal to determine the past tense of *walk* looks like:

```
PAST-TENSE-GOAL23
       ISA PAST
       OF WALK
       STEM NIL
       SUFFIX NIL
```

The goal is of type PAST (indicated by the "ISA PAST"), has the value WALK in its OF slot (WALK itself is also a declarative chunk), and has its other two slots, STEM and SUFFIX, set to NIL, indicating that they have no value yet. In order to produce a past tense, the two empty slots, STEM and SUFFIX, have to be filled. Once this goal is accomplished, the chunk is stored in declarative memory, and looks like:

```
PAST-TENSE-GOAL23
       ISA PAST
       OF WALK
       STEM WALK
       SUFFIX ED
```

As has been mentioned, the models starts out with three strategies: retrieval, analogy and the zero rule. Both retrieval and zero rule are modeled by a single production rule each.

The retrieve rule attempts to find a previous past-tense goal in memory for the word it seeks the past tense of, and, when successful, uses its STEM and SUFFIX slots to complete the current goal. The "use the stem" rule just copies the contents of the OF slot to the STEM slot and sets the SUFFIX to BLANK.

Two rules implement the analogy strategy. The strategy used is not very sophisticated: it is basically a simple pattern matcher. A first rule retrieves an example from memory and just copies the value of a filled slot from the example to the corresponding empty slot in the goal. The version of this rule that is of interest is the rule that focuses on the suffix slot:

```
RULE ANALOGY-FILL-SLOT
IF     the goal has an empty suffix slot
       AND there is an example in which the suffix has a value
THEN   set the suffix of the goal to the suffix value of the example
```

The second rule looks for a pair of slots in the example that have the same value. It then ascertains that these slots are made equal in the goal as well. The version of this rule we need for the past tense model is the rule that notices that the OF slot and the STEM slot are equal.

```
RULE ANALOGY-COPY-A-SLOT
IF     the goal has an empty stem slot and the of slot has a
       certain value
       AND in the example the values of the of and stem slots
       are equal
THEN   set the stem to the value of the of slot
```

Each of the two rules that implement the analogy strategy fills in one of the slots in the goal. The Analogy-fill-slot can retrieve two types of examples: examples with no suffix, which is the case in irregular verbs and previous experiences in which the present is used as past tense, and examples with ED as a suffix. Analogy therefore produces two types of past tenses: past tenses identical to the present tense (mimicking the zero rule), and past tenses by adding *-ed* to the stem. The former will occur much more often than the latter, since irregular past tenses are more frequent and therefore more readily available from memory.

ACT-R's production rule mechanism learns new rules by combining two rules that have fired consecutively into one. In order to restrict the number of retrievals from declarative memory to just one, the retrieval of the first rule is substituted into the rule. The resulting rule is therefore a specialization of the two parent rules. The specialization that is of particular interest occurs if Analogy-fill-slot (the rule that sets the suffix) fires first after retrieving a regular example, and Analogy-copy-a-slot (the rule that copies the stem) fires secondly. In that case the retrieved suffix (in the case *-ed* since the example is regular) is substituted into the rule itself, producing the following rule:

```
RULE LEARNED-REGULAR-RULE
IF    the goal is to find the past tense of a word and slots stem and
      suffix are empty
THEN  set the suffix slot to ED
      and set the stem slot to the word of which you want the past
      tense
```

When the retrieved example is an irregular past tense on the other hand, a less useful rule will be learned: one that sets the suffix slot to BLANK, again copying the behavior of the zero strategy.

The introduction of a new rule is restricted by two constraints. The first constraint is that a new rule is learned only when the parent rules have sufficient experience. New rules are not based on rules that have just been learned themselves. This first constraint results in the fact that it takes about a month to learn the first rule. The fact that it takes even more time to learn the actual regular rule is due to the fact that regular verbs have a low frequency, so the probability that one will be selected as an example in analogy, a strategy that itself is used only occasionally, is low.

The second constraint is that the initial evaluation of a newly learned rule is lower than the evaluation of its parents (it is derived from the evaluation of its parents, but with a penalty subtracted from it). On the other hand it receives some extra noise on this evaluation. This extra noise decays over time, but ensures the rule has some opportunity of being selected. As the noise decays, the rule needs to improve its evaluation by gaining positive experience. Fig. 2 shows some of the details: both the regular rule and the blank-suffix rule start out with an evaluation that is lower than their parent rule, analogy. The real expected gain of the regular rule, however, exceeds analogy, and the model quickly learns this is the case. As a consequence, the rule soon dominates analogy and the use-stem rule. It does not

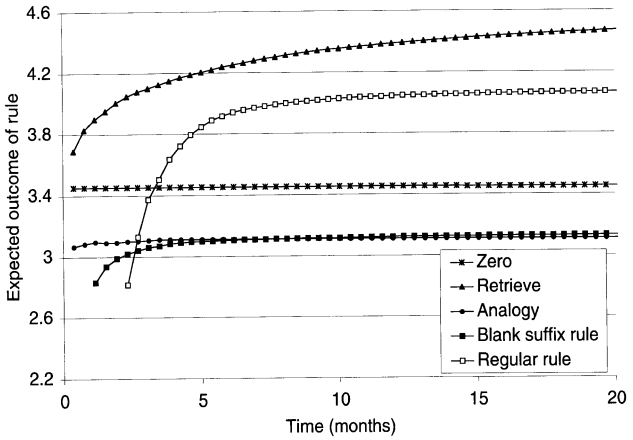

Fig. 2. Expected outcomes of the different strategies for the model. The exact scale of the expected outcomes is not relevant: it is the differences between the rules that determine the rule that wins.

surpass retrieval, however, so blocking is effectively maintained. The blank-suffix rule on the other hand does not manage to gain sufficient evaluation to counter the decaying noise, and is effectively "forgotten" by the model.

The current model uses a very limited representation of the words and implements only a very limited analogy strategy. Additional detail in these representations and strategies would lead to alternative attempts by analogy to produce past tenses, for example irregular vowel changes. In that case these attempts might be compiled into production rules as well. These rules, however, do not have the same wide applicability as the regular rule: the *ing-ang* rule based on *ring-rang* will be applicable to *bring*, but not to *work*. Before this specialized rule can participate in the rule set, it has to gain experience, but it can only gain experience on very specific examples. It will therefore be very hard for such a rule to compete with the regular and retrieval strategies that gain experience must faster.

Details on the parameters used by the model can be found in Appendix A.

4.2. Examples of how the model works

In order to gain a better understanding of how the model works, we will walk through an example of the verb *to break*, and the verb *to smash*. We will look at what the model's strategies will produce at different stages in the learning: initially, after the model has seen some examples, after the model has learned the regular rule, and finally after the model has mastered the task.

4.2.1. Initial behavior

Initially the model does not know any single example of inflecting verbs, so retrieval and analogy will always fail for both *break* and *smash*. The only viable strategy is the zero rule, which produces **break* and **smash* as past tenses, respectively.

4.2.2. Behavior after some examples have been learned

After some examples have been learned, retrieval may or may not be successful. If the right example is present in memory and it is sufficiently active, it will be retrieved, producing correct past tenses, *broke* and *smashed*. However, as *smash* is a low-frequency verb, it is unlikely to be retrieved. Analogy is now also a viable strategy, as there are examples that can be retrieved as templates. If analogy retrieves an example that does not sufficiently match *break* or *smash*, it will fail and will effectively produce **break* and **smash* as past tenses. In the current model, the analogy strategy will not find a match if the example changes the stem. It will always find a match if the example is a regular verb, or an irregular verb in which the stem is identical to the past tense (e.g. *hit-hit*). When analogy retrieves a regular example (i.e. *work-worked*), it can use this as a template to produce **breaked* and *smashed*. This will happen only very occasionally, as analogy is not a strategy that is chosen often (because it is expensive) and it will produce these only if it retrieves a regular past tense as an example, instead of the generally more active irregular past tenses. Nevertheless these occasional regularizations build up to eventually learn the regular rule.

4.2.3. Behavior after the regular rule is learned

The occasional use of the analogy strategy will at some point lead to the learning of the regular rule, as described in the previous section. Still, retrieval will remain the dominant strategy, so in most of the cases retrieval will produce the past tense, which is most of the time correct, *broke* and *smashed*. However, the pool of examples will become slightly polluted by overregularizations of irregular verbs the model itself has produced in the past (**breaked*). In general, the correct examples will be more active than the false examples, but there is no mechanism to really safeguard this. If retrieval fails to find a past tense, the regular rule is now the backup strategy, producing **breaked* and *smashed*. Analogy and the zero rule will now be used very rarely, as they now have to compete with the regular rule, and the increasingly successful retrieval strategy.

4.2.4. Behavior after the model has mastered the task

There is no clear moment at which one may judge that the model has mastered the task, but the best approximation is the moment at which all the irregular past tenses are represented as chunks in declarative memory with a sufficient and stable activation. At that stage, regular past tenses with high and moderate frequencies will also be memorized separately, making retrieval almost the exclusive strategy to use. The regular rule will only be used for low-frequency regulars and new words, as in the wug-test. At this stage the analogy and zero strategies will almost never be used anymore.

4.3. Performance of the model

The input for the model consists of the 478 words from Marcus et al. (1992). Every 2000 simulated seconds two words are presented for perception and one word is selected for generation. These words are randomly selected based on the frequency distribution in Fig. 1. The word for generation is presented to the model with the goal to find the past tense. The model receives no feedback on the accuracy of its generation. Although the model gets no external feedback, it can update the expected outcomes of its production rules based on internal feedback, caused by different execution times of the different strategies. Both the past tenses it perceives and produces are added to ACT-R's declarative memory, resulting in a growing library of past-tense examples. Not all these examples will be available, however, due to activation decay with time. Furthermore, not all examples are necessarily correct, since incorrect forms produced by the model itself are also maintained.

The repeated application of analogy results in the learning of two new production rules. One of these is a production rule that just uses the stem. This is identical to the zero rule and plays no significant role in the behavior of the system. The other learned production rule uses the *-ed* inflection, the rule that will produce the overgeneralizations. If we had a richer phonological system we might also learn production rules for the few other semi-regular vowel-change patterns in English like *ring-rang*. However, errors produced by such rules are such a small part of the English past-tense system (Xu & Pinker, 1995) that their omission does not change our ability to capture the basic U-shaped learning curve. The fact that these irregular vowel changes are rare is consistent with the model, since a

successful analogy can only be made if a suitable example (e.g. *ring* while processing *bring*) is retrieved.

The analogy strategy is not very successful in general, because the example it retrieves is the verb with the highest activation. These examples are irregular most of the time, so only when a regular example is retrieved will the analogy strategy produce a regular past tense. Together with the fact that the proceduralization mechanism requires that the parent production rules (in this case the rules that implement the analogy strategy) have sufficient experience, it will take some time to learn the regular rule. The regular rule, once learned, is very successful: it will always produce a past tense given any stem. Only retrieving a past tense from memory is more efficient. Once the new production rule is learned, it takes some more time before it is used frequently, because its expected outcome still has to grow.

Fig. 2 shows how the expected outcomes of the different strategies develop in the first 20 months of the simulation. Note that production rule selection is a stochastic process: the rule with the highest evaluation only has the highest probability of being selected. Furthermore, if a certain production rule fails (e.g. the retrieval rule cannot retrieve anything) the next best rule is selected (e.g. the regular rule). In each month approximately 1300 past tenses are produced. This number is chosen somewhat arbitrarily, but the model is not critically dependent on the exact rate of production. Initially, the model will attempt to retrieve a past tense from memory, and, if this fails, it will use the stem as the past tense. As it gains experience in storing past tenses from the environment and producing them itself, however, it will be more and more successful in retrieving previous past tenses from memory. The expected outcome of the retrieval production rule increases gradually over time as the model learns more past tenses. The two other initial strategies have a stable evaluation over time, since they are not influenced by the learning process.

Around the first month of the simulation the analogy strategy has gained enough experience to allow specialization, so the model learns its first new production rule, the rule that uses the stem as the past tense. As this rule duplicates the behavior of an already existing rule, the zero rule, it does not play an important role in behavior. In the second month the model also learns the production rule corresponding to the regular rule. Its expected outcome initially increases rapidly, because it is able to produce a past tense out of any stem, as opposed to the analogy strategy that it originated from. Around the fourth month of the simulation the evaluation of the regular rule passes the evaluation of the zero rule. This changes the basic behavior of the model to first try to retrieve an irregular past tense from memory, and if this fails to create a regular past tense by adding the *-ed* suffix. The reason why retrieval remains the dominant strategy is the fact that it has no phonetic post-processing costs involved with it. The retrieval production rule can retrieve both regular and irregular past tenses. As learning progresses, regular verbs become more stable in declarative memory, and retrieval can handle both regular and irregular inflection. Just as retrieval gradually becomes an alternate way to handle regular words it eventually provides a basis for handling many relatively low-frequency irregular verbs. Only for the very infrequent verbs or novel words does the regular rule remain as the dominant basis for generating past tenses.

Fig. 3a shows the proportion of responses the model produces over time. The proportions of correct responses for both regular and irregular verbs are shown, as well as the

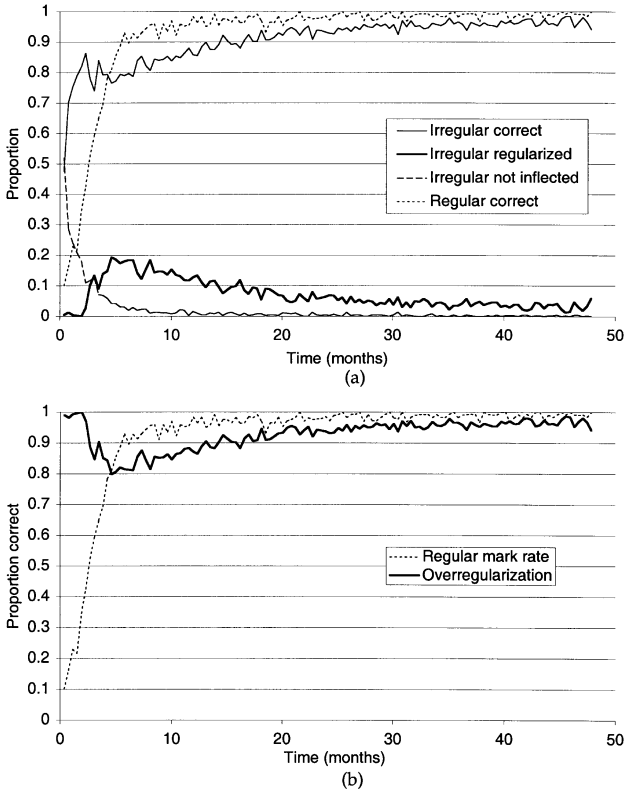

Fig. 3. Results of the model. (a) Proportions of responses by the model over time. Incorrect regulars are not indicated since these are all "Regular not inflected". (b) Overregularization of the model as it is usually plotted: overregularization is equal to (irregular correct)/(irregular correct + irregular regularized), and regular mark rate equals (regular correct)/(regular correct + regular incorrect).

errors made with irregular verbs, which can be either an error of non-inflection or over-regularization. All errors on regular verbs are errors of non-inflection; therefore, since the proportion of errors is one minus the proportion of correct inflections, it is omitted from the graph. It is usually hard to detect errors of non-inflection in actual data, because only in contexts where it is clear that a past tense should have been used (like in "Yesterday Daddy go…") can the error be recognized. The data are usually plotted like in Fig. 3b, where overregularization equals the number of correct responses on irregular verbs divided by the sum of correct irregulars and irregulars inflected regularly.

The results show U-shaped learning, at least when they are plotted in the same way as the data usually are (Fig. 3b). The downward slope coincides with the learning of the regular rule. At this point in the simulation the model has not memorized all irregular past tenses yet at a level that they can be retrieved without errors. If it fails to retrieve an irregular past tense it will use one of the regular rules, producing overregularization. The regular rules may also win the competition with the retrieve production rule because of the stochastic noise, so the model will not even try to retrieve an irregular past tense. A third

source of overgeneralization occurs if the retrieve production rule retrieves a previous overgeneralization from memory. Gradually the model will master the irregular past tense, producing the upwards slope in the U-curve. Contrary to neural networks models, there is no corresponding U-shape in regular past tense behavior, even the opposite: at the onset of the U-shape in irregular verbs, performance on regular verbs increases dramatically.

An interesting observation about Fig. 3a is that accuracy shows hardly any U-curve at all. The main change visible there is a change in the type of errors from non-inflection to overregularization. As errors of non-inflection are much harder to detect than overregularization in the case of children, there seems to be an increase in errors to the outside world, while actually none are present.

The model treats equivalently input from the environment and examples that it produces. When past tenses from the outside world concur with past tenses produced by the system itself, they strengthen each other. If we assume forms from the outside world are always correct and forms produced by the model are occasionally correct, incorrect past tenses in declarative memory eventually lose the competition.

The predictions that model produces can be compared to empirical data from Marcus et al. (1992). From the children that they studied most extensively, Adam and Sarah show the pattern associated with U-shaped learning, that is, they show a reliable period without overregularization followed by a period with overregularization. A third child, Abe, shows extensive overregularization over the whole period that he is studied. He shows no signs of U-shaped learning, presumably because his overgeneralization had already started at the beginning of the study. He is of particular interest because Marcus et al. report on his behavior on individual words. We will look at how the model handles individual words later on.

Fig. 4 shows the overregularization rates of Adam and Sarah. Both the children and the model show the initial stages of U-shaped learning, from no overregularization (first stage) to overregularization (second stage). Although overregularization in the model gradually diminishes, Adam and Sarah do not show any signs of diminished overregularization during the period studied. Nevertheless Marcus (1996) reports that overregularization gradually trails off in children, 4.2% in preschoolers, 2.5% in first graders and 1% in fourth graders. Even adults sometimes overregularize, but this is very rare. In both Adam and Sarah overgeneralization seems to increase more gradually than in the model. A possible explanation for this is the fact that the model is tested on the full vocabulary, even words it has not yet encountered very often. An irregular verb that the model has not yet encountered will almost always lead to an overgeneralization. Children on the other hand will presumably tend to avoid words they do not know well. As their vocabulary grows they will tend to use less frequent irregulars more often, increasing overgeneralization.

In the children, the best predictor for the onset of overregularization is a sudden increase in the rate in which regular verbs are actually marked for past tense, as is indicated by the dotted lines in Fig. 4 (the sudden spike between months 27 and 32 in Sarah's graph is due to the small number of observations). This increase indicates the discovery of the regular rule. Fig. 3 shows that this is true in the model as well: again the dotted line indicates the rate in which regular verbs are marked for past tense.

Another aspect of children learning the past tense is individual differences. Adam

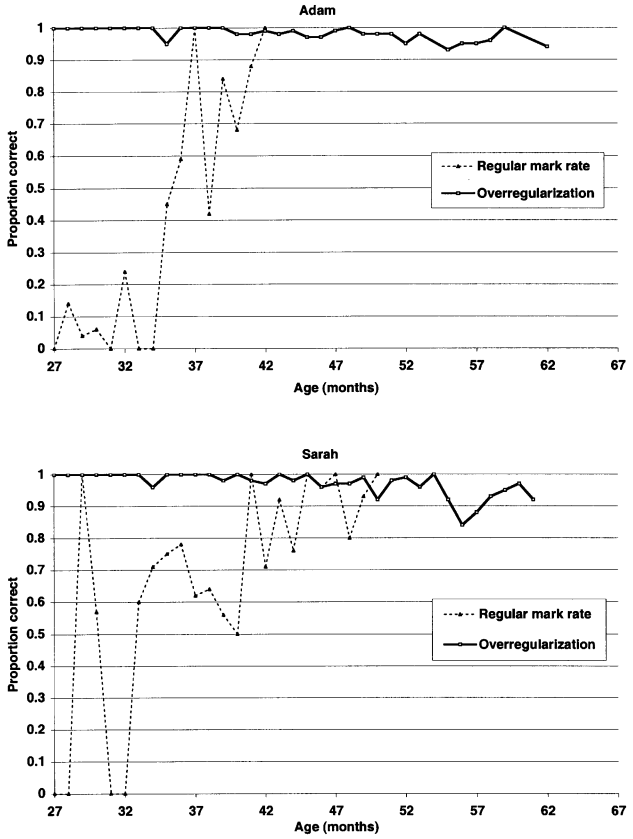

Fig. 4. Overregularization and proportion of regular past tenses marked by Adam and Sarah. Adapted from Marcus et al. (1992).

exhibits very little overregularization, while Sarah shows much more tendency to overregularize. A possible explanation can be found in the input from the environment. In the current model the two examples are perceived in the environment for every example produced. If fewer examples are perceived, because the environment supplies fewer or the child pays less attention to them, overregularization will increase. Fig. 5 shows overregularization for different ratios of examples perceived and produced.

Although the vocabulary that serves as input for the model is fixed, the model itself only acquires these words over time. A good estimate of whether or not a certain word is part of the vocabulary is to look at its activation. If this activation is past a certain threshold, the word is assumed to be part of the vocabulary of the model. Fig. 6 shows the result of the model, together with data from Adam and Sarah. As can be seen, both for the model and the children there is a gradual increase in both regulars and irregulars. There is no sudden spurt in vocabulary for either regulars or irregulars.

Marcus et al. (1992) also report behavior on individual words by Abe. The left side of Fig. 7 shows the four examples, and the right side words from the model. For the word *Say*

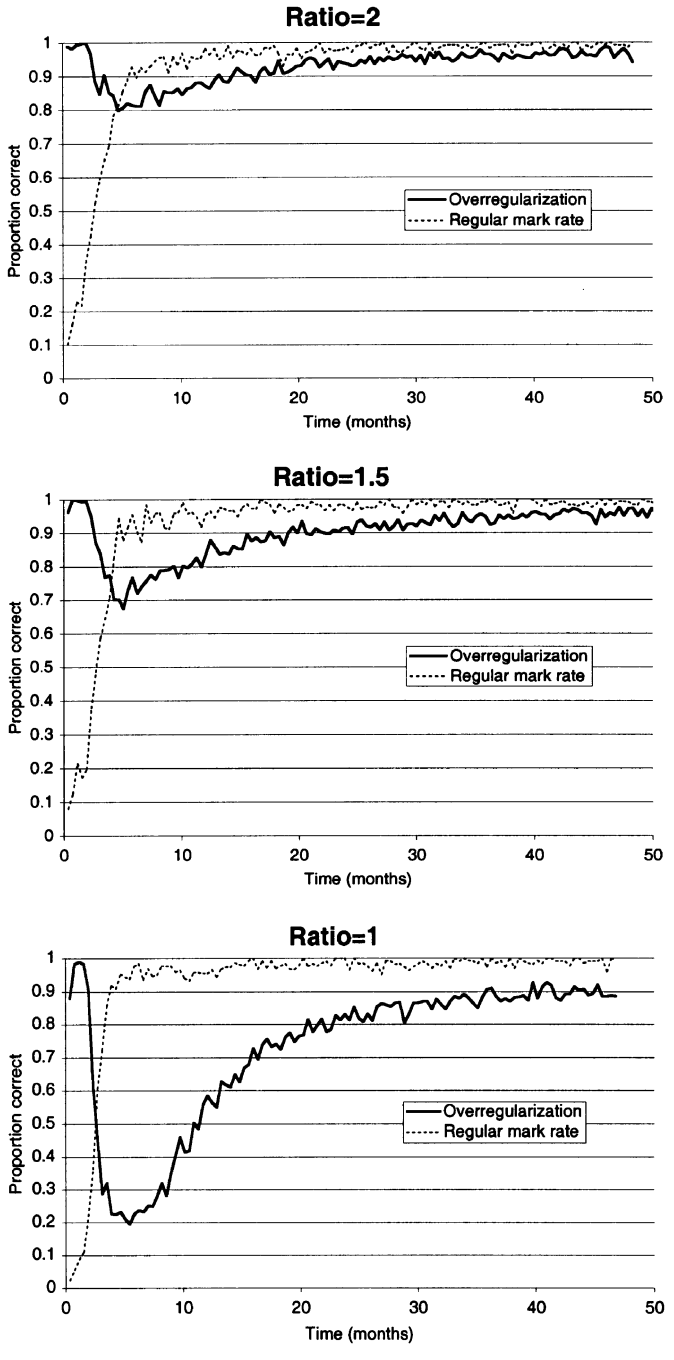

Fig. 5. Overregularization for the model for different ratios of input from the environment and production.

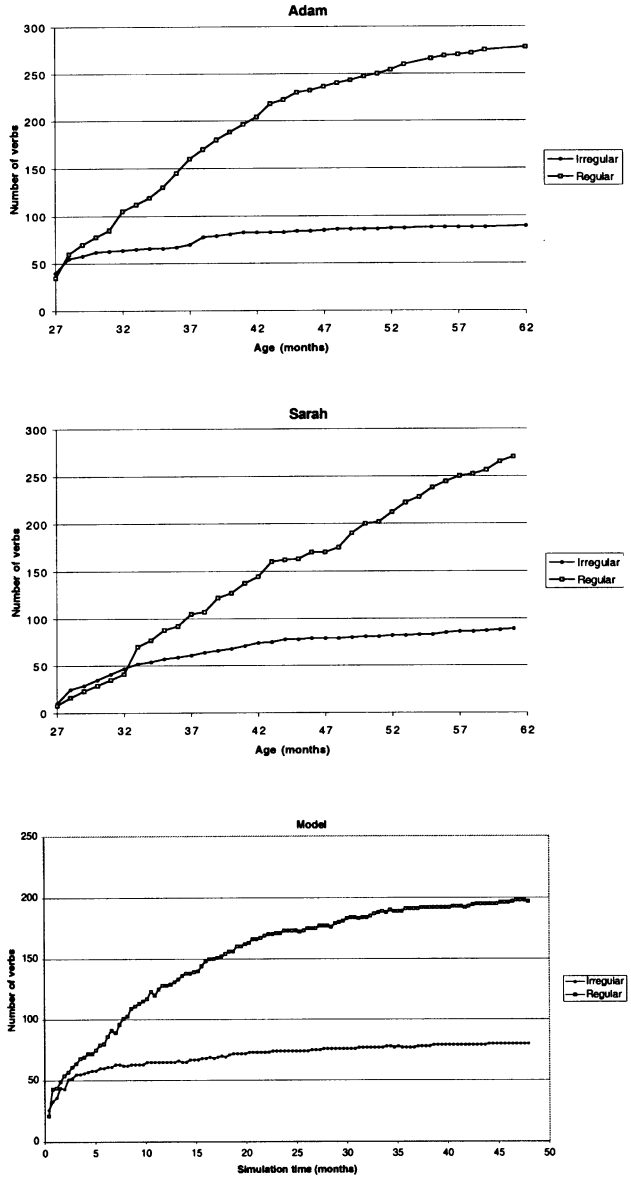

Fig. 6. Vocabulary growth for Adam, Sarah (adapted from Marcus et al., 1992) and the model.

Abe shows very little overgeneralization. This turns out to be true for the model as well, and can be explained by the fact that *Say* is a high-frequency irregular verb (no. 4 in the word-frequency list). The second example, *Eat*, is somewhat less frequent (no. 114). Abe has some early problems with this word, but recovers later on. The model also needs more time to master *eat*. Note that the curve for the model is much smoother than Abe's. This

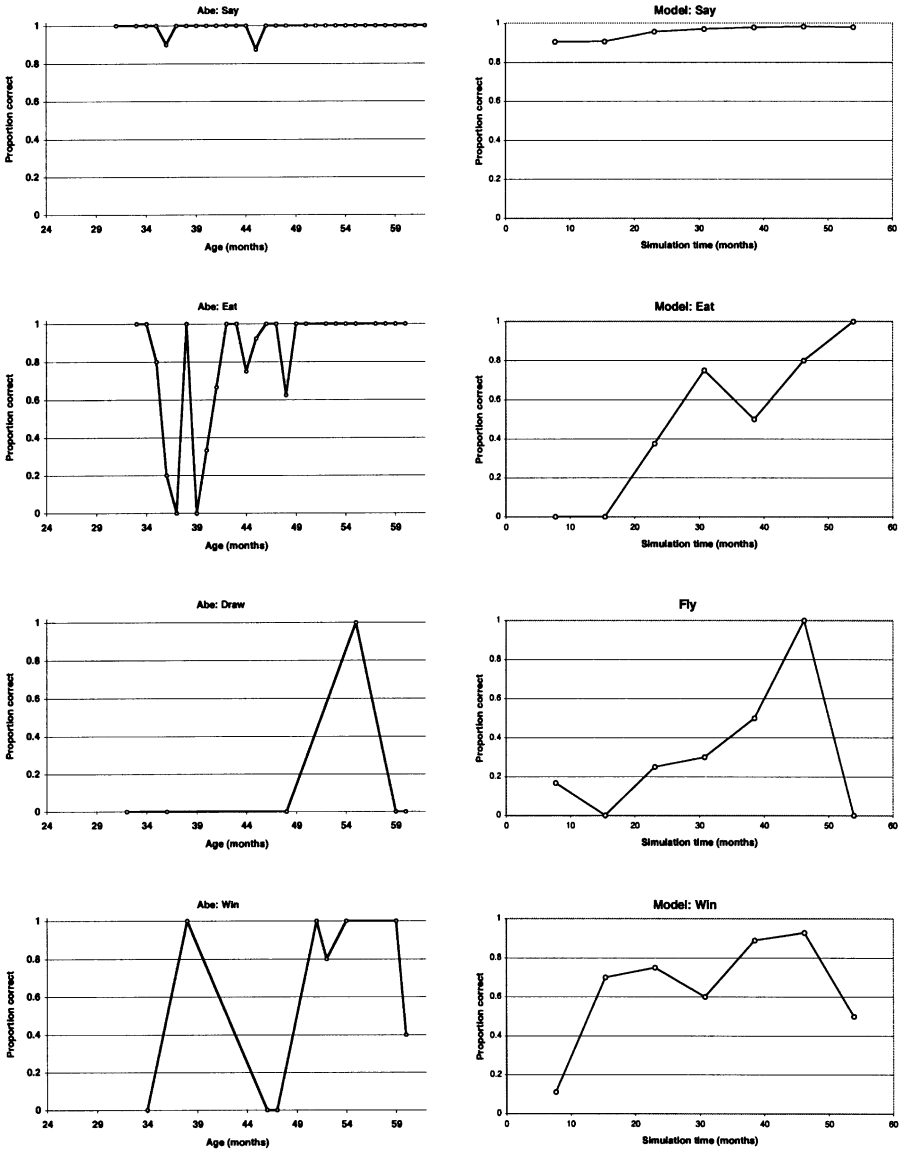

Fig. 7. Overregularization of individual words for Abe (adapted from Marcus et al., 1992) and the model.

may be attributed to the larger number of observations. For irregular words that Abe uses very little, his behavior is rather erratic, as exemplified by the words *Draw* and *Win*. This is true for the model as well (we took *fly* to match *draw*, because *draw* is actually fairly high-frequent, no. 73). The erratic patterns in low-frequency words can be attributed to the stochastics sampling procedure. These results concur with a study carried out by Stem-

berger and MacWhinney (1986), who found that inflection errors are much more frequent in low-frequency irregulars than in high-frequency irregulars.

In general, overregularization is most common for infrequent words. In the model this can be explained by the fact that low-frequency irregular forms have a lower activation, increasing the probability of a retrieval failure and subsequent regularization. Marcus (1996) reports that this is indeed the case: he found a correlation of -0.34 between the frequency that parents use an irregular verb and the children's overregularization rate. This correlation is -0.59 for the model (using the Francis and Kucera (1982) frequencies for the parental frequencies). This correlation is probably larger because the model is trained on and compared to the exact Francis and Kucera frequencies, while these are somewhat noisy estimates in the Marcus analysis.

4.4. Low-frequency regulars

The German plural shows that the default rule is not necessarily the dominant inflection. In order to investigate whether our model was capable of capturing the right rule, even if regular forms are of low-frequency, we took the same model but modified the input vocabulary, so that the type-frequency of regular verbs is below 10% (46 words), and the token-frequency below 5%. The model was run using exactly the same parameters. Fig. 8 shows the results of this simulation. Interestingly enough, the results are almost identical to the normal simulation, with the exception that the onset of overregularization is slightly later in the new simulation. The reason why the frequency of the default rule does not significantly affect the learning of the rule is that the rule, once learned, turns out to be a very effective rule.

Note that this model only demonstrates how a low-frequency default rule can be learned. The German plural itself is more complicated, as there are several different suffixes that compete as default rules. All of the non-default suffixes, however, are bound by specific linguistic constraints that limit their general applicability. Taatgen (2001) has made a more elaborate model of the competition among different default rules using the same type of model described here. Also note, as has already been remarked in Section 1, that the status of the default rule in German is still under debate (Hahn & Nakisa, 2000).

5. Comparison between neural network models and the ACT-R model

We will compare our model to the Plunkett and Marchman (1993) and Plunkett and Juola (1999) models, mainly because they explicitly model the U-shaped learning that our model focuses on (not all models do), but the comparisons may largely extend to other neural networks as well. Fig. 9 gives an overview of the processing assumptions made by both neural network models and the ACT-R model. Parts (a) and (b) of that figure represent the processes of perception and generation in the neural network models and parts (c) and (d) represent the corresponding processes in ACT-R. Both models only partially model all processes involved in interpreting and generating the past tense. The processes not modeled but assumed are indicated by dashed boxes. Both models focus on generating past tenses from stems. The reverse process, generating the stem given a past tense, is not

modeled in either model. But both models assume that the results of this reverse transformation are available to the model. A difference in this aspect is that although the past-tense-to-stem transformation is not explicitly modeled in ACT-R, the declarative knowledge can be shared between perception and generation. The neural network model has to train a separate network to accomplish the task. Another advantage of the sharing of information is that the perception of the past-tense part of the process is much more simple and natural in ACT-R (part c) than neural models (part a). Whereas storing the example of the past tense in declarative memory is an automatic by-product of interpreting it, the neural network model has to make additional assumptions about a hypothesis generator that is not a necessary part of the interpretation process. An even more implausible part of this process is the input/uptake filter. A property of words in the real world is that some are really low-frequent. Also, the spacing between occurrences is not necessarily even: the same word may be used a number of times in a row, and then not turn up for several weeks. The input/uptake filter therefore has to do a number of jobs. It has to take care that the

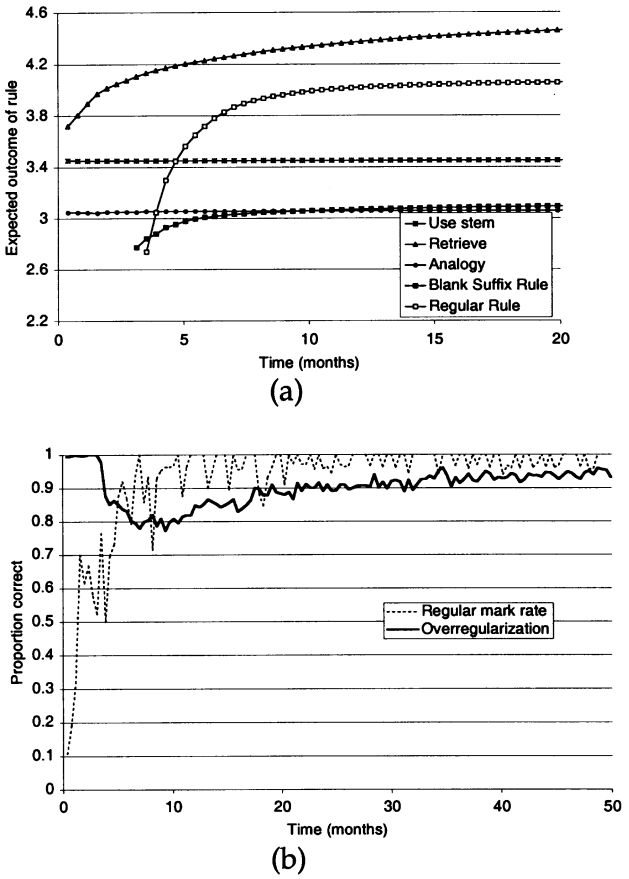

Fig. 8. Results of the German plural simulation: (a) expected gains; (b) overregularization.

vocabulary increases gradually, but often not according to the real growth curves like in Fig. 6, but rather a discontinuous or exponential curve. It also has to make sure low-frequency verbs occur often enough, and high-frequency words do not occur too often. Filtering out high-frequency words is a doable job, but it is not clear whether low-frequency verbs occur often enough in the input at all. The ACT-R model needs no such assumptions: all past tenses, whether previously known or not, are stored in declarative memory directly. The low-frequency regular run of the ACT-R model shows that it has the potential of capturing the German plural as well. It is doubtful whether neural network models can come up with an input/uptake account that is able to learn the regular rule in this case, and that still is plausible.

Both models also differ with respect to production. Where the neural network is passive in its production, with no learning going on, production is the main source of learning in

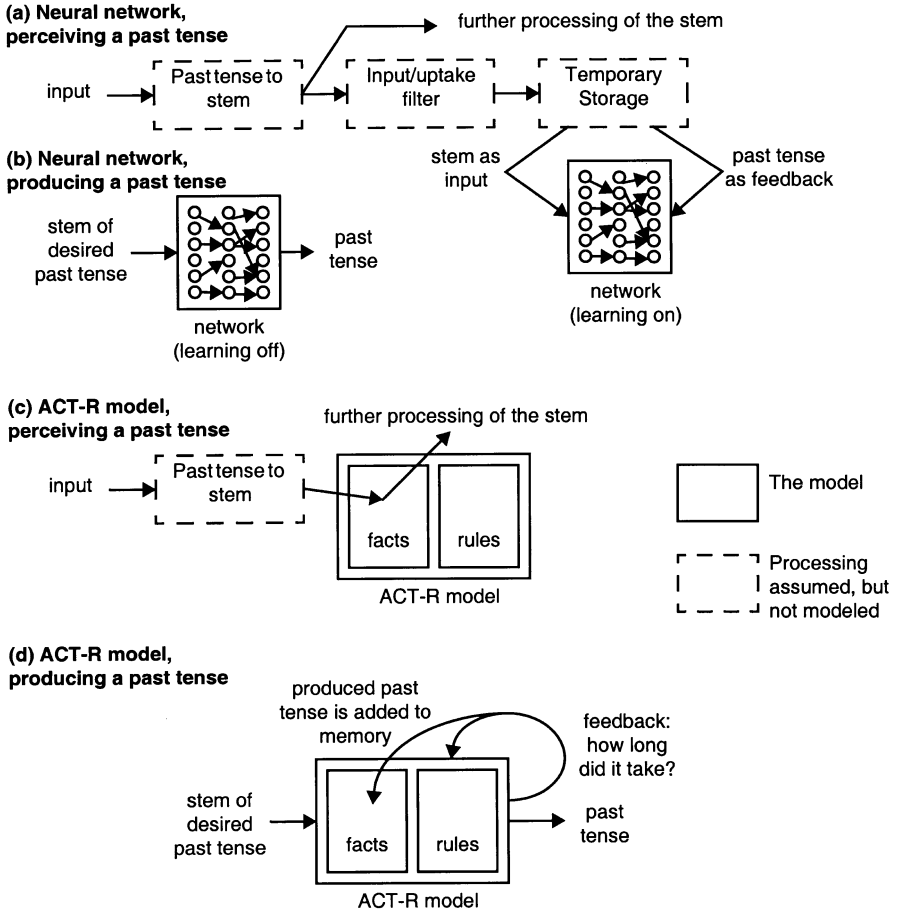

Fig. 9. Overview of the processing assumptions in neural network models and the ACT-R model. Dotted boxes are aspects of processing that are assumed, but not explicitly modeled.

the ACT-R model: it uses its internal feedback to adjust its strategies, and stores its own produced past tenses along with examples it has perceived. In learning in general, practice is a main component of the learning process.

Another issue mentioned in Section 1 is the critical mass hypothesis, on which neural network models rely. The ACT-R model is sensitive to the frequency of the regulars but does not depend on a critical mass of regulars. The learning of the regular rule relies on practice with the analogy strategy and sufficiently activated regulars. Both these conditions require a reasonably sized and activated vocabulary. As a consequence, the low-frequency regular version of the model is later in learning the regular rule, because it takes more time to properly learn regular examples and use them in the analogy strategy.

6. A model without any input

An interesting variation on the model presented here is to deprive it of all input, in order to see what happens in a minimal system. Since the model has no access at all to correct past tenses, it will have to invent them itself. As a consequence, it not only needs strategies that are usually associated with past-tense learning, but also some production rules that make up new past tenses. Hare and Elman (1995) have studied a variant of this situation. They started by training a network on past tenses in Early Old English. This network was used as a teacher for a new, generation 1, network. Consequently, the generation 1 network was used to train generation 2, up to generation 5. The result was that many low-frequency irregular verbs became regular verbs, a development that can also be observed in English itself. The model was, however, not able to come up with new irregular verbs, something that does happen in real language.

In our ACT-R model without input we added a production rule to generate new past tenses. This rule can do one of two things: it can modify the stem into something new (corresponding to an irregular past tense), or it can add a random suffix to the stem (corresponding to a regular past tense). Since several suffixes are generated by the model, there is a potential for multiple regular rules with different suffixes. Generating something new is an expensive strategy, so it has a high cost associated with it. The model will therefore prefer the other strategies as soon as they are reasonably productive. Another change in the model is that regular past tenses now have a penalty because they are longer, a fact that we mentioned earlier, but on which the previous model did not rely.

The simulation is set up similarly to the previous model: each 1000 seconds the model has to produce the past tense of a given word, and the model is run for 60 simulated months. Fig. 10 shows the proportion of times the model uses an irregular or a regular past tense for a certain word. It clearly prefers irregular forms for high-frequency words and regular forms for low-frequency words, as is observed in natural language. Due to the probabilistic nature of both the model and the selection of words from the vocabulary, there are many spikes in the graph, indicating infrequent words for which the model nevertheless prefers an irregular past tense and vice versa. Note that this particular model has no incentive to use the same past tense all the time, so it can use different forms of past tense without penalty.

Although this model receives no feedback at all, it is still interesting to look at the

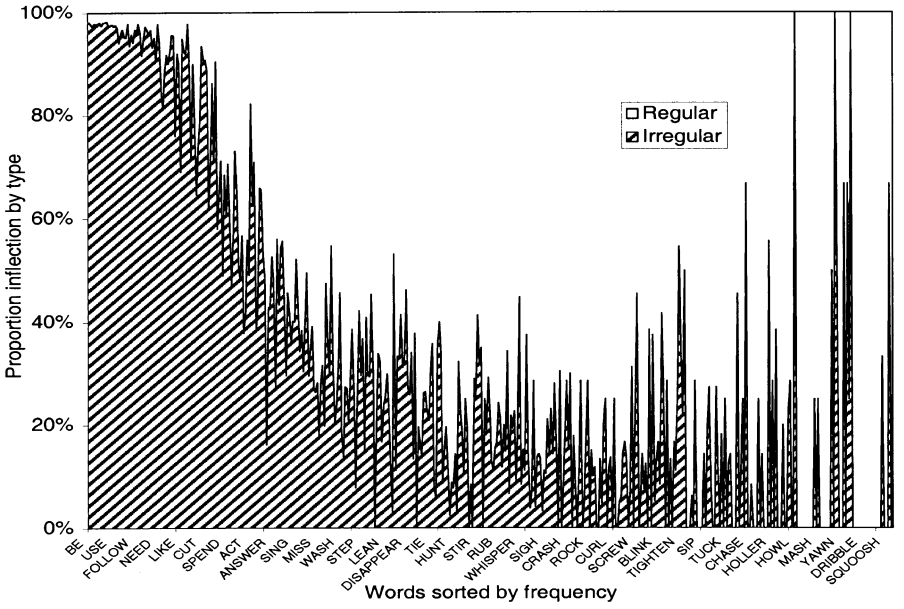

Fig. 10. Proportion of times the model chooses an irregular or a regular past tense for a certain word. Words on the *x*-axis are sorted by frequency as in Fig. 1.

tendency of the model to use irregular forms for verbs that are irregular in reality. The model has no way of knowing what words are regular or irregular, but irregular words generally have a high frequency, so the model will eventually favor irregular past tenses for these words. Fig. 11 plots how often the model chooses an irregular form for an irregular verb as opposed to a regular form. Two different runs are shown, because this model exhibits more randomness than the previous one. Interestingly enough, the model shows U-shaped learning. The run at the top of Fig. 11 actually shows three U-shapes, each corresponding to the learning of a different regular rule. In the bottom run in Fig. 11 regular rules are acquired earlier, at a time when still few past tenses are memorized. Therefore, the rules have a larger impact on performance, as exhibited by a deeper U-curve.

While the no-feedback model does not represent the learning situation of the typical child, it nonetheless illustrates in a particularly clear way the forces driving the inflectional development of the model. It shows that without feedback, without any change in vocabulary, and without an empirically incorrect high token-frequency of regular words, one will tend to learn to inflect high-frequency words irregularly and low-frequency words regularly. Moreover, the U-shaped learning is caused by a model that only tries to learn production rules and memorize examples at the same time. The initial appearance of overregularization simply reflects the greater scope of the regular rule while the reduction of overregularization reflects the greater efficiency of the irregular construction.

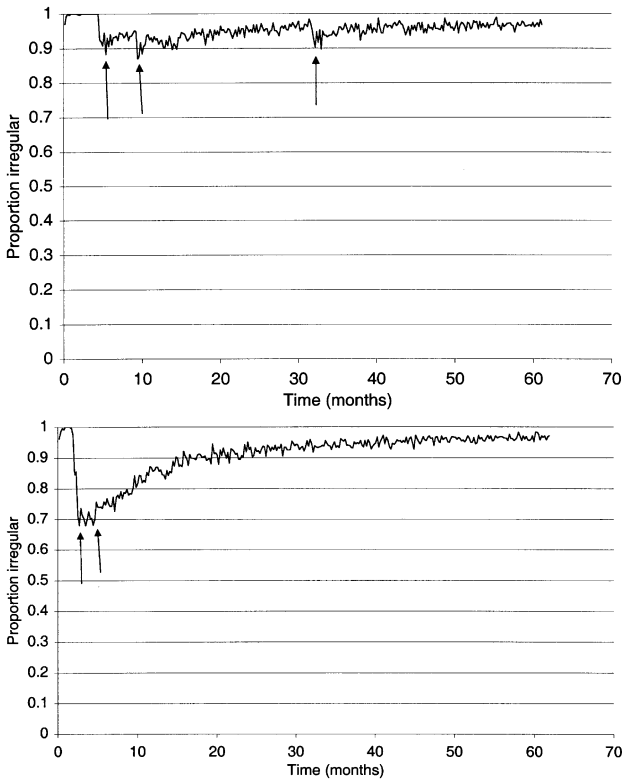

Fig. 11. Proportion of times the no-feedback model produces an irregular past tense for an irregular verb as opposed to producing a regular past tense. Two different runs of the model are depicted. The moments new regular rules are learned are indicated by arrows.

7. Conclusions

The basic hypothesis underlying the models presented in this paper is that regular and irregular forms each have their own advantages. Irregular forms are more effective as long as their use is frequent enough, while regular forms always work at a slightly higher cost. Concurrent with this trade-off is the trade-off between retrieval and the use of a rule. Retrieval is more efficient, since using a rule requires phonetic post-processing, but retrieval is possible only if an example is available from memory. Each of these trade-offs can explain much of the data associated with the learning of past tenses, without the need for further assumptions about the ratio between regulars and irregulars, feedback or the growth of the vocabulary. Instead the model is capable of generating the growth of vocabulary instead of prerequiring it. The models discussed in this paper do not model words at the level of phonetics. We believe that for this model, its simplicity is a virtue because it makes more transparent the critical forces at work. The basic phenomena of learning irregular and regular past tenses can be explained by the small set of principles on which the model is based. It is however interesting to observe that words ending in a *-d* or

-t are more often irregular (35% of the irregulars end with a *-d* or *-t*, as opposed to 11% of the regulars). This can be explained by the fact that for these words the additional costs of regularization are higher than other words, since the *-ed* has to be pronounced as a separate syllable instead of just an added phoneme.

The fact that all three models (majority default, minority default and no input) exhibit similar behavior, despite their differences, may raise the criticism that the behavior is purely produced by the mechanics of the model. Instead, the behavior of the model depends on the characteristics of the input (or, in the no input model, the characteristics of the forms it is allowed to generate). If, in the language, there is a systematic way to inflect a word without incurring any penalties (phonetic or pronunciation), this rule would dominate behavior. Also, if no pattern of inflection can be found that can be applied to all words, the use of a rule will be limited or absent. To summarize, the fact that the model learns a rule is not because it is an innate process, but because the combination of the structure of the environment and the striving of the cognitive system for efficiency (both time and memory) makes the learning of a rule the best solution.

The views expressed in this model largely coincide with the dual-representation account as offered by Pinker and Marcus. It can fill in some of the gaps in this account. It shows how the regular rule can be learned by analogy. It explains rather than assumes that blocking, implemented by the retrieval strategy, dominates the regular-rule strategy. It offers an explanation for why it takes so long to find the right balance between examples and the regular rule. The main reason is the lack of feedback. But another reason is the fact that the regular rule serves an important function at the stage in learning where the child has only memorized a few of the irregular past tenses. It allows a child to communicate the past tense reasonably efficiently although it has mastered only a fraction of the vocabulary. A third reason is the fact that children learn their own errors (see also Platt & MacWhinney, 1983). Correct irregular past tenses have to compete in memory with overregularizations. This offers an explanation for the phenomenon that correcting a child doesn't seem to help, exemplified by the following exchange reported by Cazden (1972).

Child: My teacher holded the baby rabbits and we patted them.
Adult: Did you say your teacher held the baby rabbits?
Child: Yes.
Adult: What did you say she did?
Child: She holded the baby rabbits and we patted them.
Adult: Did you say she held them tightly?
Child: No, she holded them loosely.

In this exchange the number of past tenses produced by the child equals the perceived examples from the environment, so both the wrong and the correct form are strengthened equally. Neural network models will have trouble accounting for this, as no learning occurs during production, and continued perception of the correct example should reinforce the right behavior quickly.

The penalty for overregularization is very small: it is slightly less efficient than using the proper irregular form and it takes time to fully exploit this difference. Pinker (1999) offers a related view on irregular words. According to his account, irregular words stem from

earlier versions of the language and survive because they have a high frequency, as opposed to low-frequency words that are regularized. As Pinker acknowledges, this cannot explain why new irregular verbs enter a language (e.g. *dive-dove*, *catch-caught*). If irregular verbs would be a historical matter entirely, they would all gradually disappear.

In general people seem to strive for short-cuts in language as long as this does not lead to communication problems. The first thing a novice in a new organization has to learn is all the acronyms his or her new colleagues use to refer to the various departments and services in the company (e.g. the Groningen University uses "DOOP" to refer to the Department of Research, Education and Planning). Within the company all these acronyms are high-frequency words. The novice, however, will initially suffer from a version of overgeneralization by using the full title of the department as opposed to the acronym. Another recent abbreviation is using *gonna* for *going to* and *wanna* for *want to*. According to Tabor (1994), the frequency of use of the verb *want* has increased in the past few centuries, especially in conjunction with *to*. Although these contractions are heavily constrained by syntax, the increased frequency of the combination has made memorizing the abbreviation at least a potential solution.

The example of *want to* shows the potential and the limitations of the model. It cannot explain how syntactic constraints can be satisfied to make contraction possible at all, but it can explain that contraction may become a viable option as the frequency of use increases. As the model is implemented in a general cognitive architecture, constraints due to phonetics, syntax or semantics are not captured as long as a general phonetic, syntactic or semantic framework isn't part of the model. U-shaped learning on the other hand can be explained by general principles of information processing and learning. Recent work by Misker and Anderson (2002) shows that ACT-R is capable of modeling aspects of language acquisition based on linguistic constraints, in their case constraints provided by Optimality Theory (Prince & Smolensky, 1993).

We believe that we have presented a new model of the past tense that has some interesting advantages over existing models. Although it uses two types of representation, its overall processing schema is much more parsimonious than the schema neural networks use (Fig. 9). All the ingredients that are used for the model, among which are the two memory systems, are parts of the ACT-R architecture, which has been validated through extensive separate experiments, or part of the ACT-R modeling tradition, like instance-based learning and the use of analogy. The model can explain U-shaped learning in different situations, and uses a feedback-schema that, when extended to other situations, may help to shed more light on the general question of the learnability of language.

Acknowledgements

This research was supported by a NATO-Science Fellowship awarded to Niels Taatgen by the Netherlands Organization for Scientific Research (NWO) and by ONR grant N0014-96-I-0491. The authors would like to thank Brian MacWhinney, Jay McClelland, Gary Marcus and Steven Pinker for their comments on an earlier draft of the paper.

Appendix A

The models are available on the internet by following the "published models" link on the ACT-R webpage: http://act.psy.cmu.edu/.

The model uses the following parameters (see Table 1): $F = 0.5$, $f = 0.25$, $tau = 0.3$, noise on expected outcome is 0.085, $d = 0.4$, $G = 5$, $W = 0.5$ (this parameter indexes the contribution of the context to the activation), noise on activation is 1.6 but decreases during the run for individual chunks, proceduralization occurs after 25 experiences with the parents rules, optimized learning is on. The phonetic post-processing cost is set to 0.6 (this cost is assigned to the analogy strategy, but is inherited by any rules that are specialized on the basis of analogy), and the cost of non-inflection (use of the present tense as the past tense, so making it necessary to indicate past tense by some other means like "yesterday") is set to 0.9. The following learning mechanisms are switched on: base-level learning, production parameter learning and production learning. Associative learning was switched off because it didn't contribute anything to the model and only slowed it down, while strength learning was switched off because it is obsolete in the new production learning mechanism. For the purposes of Fig. 6, a threshold of -3.5 was used to determine whether or not a word was part of the vocabulary of the model.

Although Figs. 2 and 3 show just a single run of the model, runs do not differ very much from each other, the main difference being the moment the regular rule is discovered. Fig. 5 gives some impression of the variance.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Burzio, L. (1999). *Missing players: phonology and the past-tense debate*. Unpublished manuscript, Johns Hopkins University at Baltimore, MD.

Cazden, C. B. (1972). *Child language and education*. New York: Holt, Rinehart and Winston.

Clahsen, H., Rotweiler, M., Woest, A. & Marcus, G.F. (1992). Regular and irregular inflection in the acquisition of German noun plurals. *Cognition*, *45*, 225–255.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage: lexicon and grammar*, Boston, MA: Houghton Mifflin.

Hahn, U., & Nakisa, R. C. (2000). German inflection: single route or dual route? *Cognitive Psychology*, *41*, 313–360.

Hare, M., & Elman, J. L. (1995). Learning and morphological change. *Cognition*, *56*, 61–98.

Lebiere, C., Wallach, D., & Taatgen, N. A. (1998). Implicit and explicit learning in ACT-R. In F. Ritter & R. Young (Eds.), *Proceedings of the second European conference on cognitive modelling*. Nottingham: Nottingham University Press.

Ling, C. X., & Marinov, M. (1993). Answering the connectionist challenge: a symbolic model of learning the past tenses of English verbs. *Cognition*, *49* (3), 235–290.

Logan, G. D. (1988). Towards an instance theory of automization. *Psychological Review*, *22*, 1–35.

MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development*, *43* (1), 1–123.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: revising the verb learning model. *Cognition*, *40*, 121–157.

Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: a test of the critical mass hypothesis. *Journal of Child Language*, *21*, 339–366.

Marcus, G. F. (1995). The acquisition of the English past tense in children and multilayered connectionist networks. *Cognition*, *56*, 271–279.

Marcus, G. F. (1996). Why do children say "Breaked"? *Current Directions in Psychological Science*, *5*, 81–85.

Marcus, G. F., Brinkman, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: the exception that proves the rule. *Cognitive Psychology*, *29*, 189–256.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57* (4), 1–182.

Misker, J. M. V., & Anderson, J. R. (2002). Combining optimality theory and ACT-R, submitted for publication.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S. (1999). *Words and rules: the ingredients of language*. New York: Basic Books.

Pinker, S., & Prince, A. (1988). On language and connectionism: analysis of a distributed processing model of language acquisition. *Cognition*, *28*, 73–193.

Platt, C. B., & MacWhinney, B. (1983). Error assimilation as a mechanism in language learning. *Journal of Child Language*, *10*, 401–414.

Plunkett, K., & Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science*, *23*, 463–490.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: implications for child language acquisition. *Cognition*, *38*, 43–102.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, *48*, 21–69.

Plunkett, K., & Marchman, V. (1996). Learning from a connectionist model of the acquisition of the English past tense. *Cognition*, *61*, 299–308.

Plunkett, K., & Nakisa, R. C. (1997). A connectionist model of the Arabic plural system. *Language and Cognitive Processes*, *12*, 807–836.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8* (1), 1–56.

Prince, A., & Smolensky, P. (1993, April). Optimality theory: constraint interaction in generative grammar. Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: explorations in the microstructure of cognition* (pp. 216–271). Cambridge, MA: MIT Press.

Salvucci, D. D., & Anderson, J. R. (1998). Analogy. In J. R. Anderson & C. Lebiere (Eds.), *The atomic components of thought* (pp. 343–383). Mahwah, NJ: Erlbaum.

Stemberger, J. P., & MacWhinney, B. (1986). Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition*, *14* (1), 17–26.

Taatgen, N. A. (2001). Extending the past-tense debate: a model of the German plural. *Proceedings of the twenty-third annual conference of the Cognitive Science Society* (pp. 1018–1023). Mahwah, NJ: Erlbaum.

Tabor, W. (1994). *Syntactic innovation: a connectionist model*. PhD dissertation, Stanford University, Stanford, CA.

Xu, F., & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, *22*, 531–556.